SIQ: easy quantitative measurement of mutation profiles in sequencing data

Robin van Schendel[1,]*, Joost Schimmel[1] and Marcel Tijsterman[1,2,]*

[1] Human Genetics, Leiden University Medical Center, Leiden, the Netherlands

[2] Institute of Biology Leiden, Leiden University, Leiden, The Netherlands

* To whom correspondence should be addressed. Tel: +31 71 5269609; Email: R.van_Schendel@lumc.nl, M.Tijsterman@lumc.nl

ABSTRACT

Establishing mutational outcomes after genome editing is of increasing importance with the advent of highly efficient genome-targeting tools. Next-generation sequencing (NGS) has become a vital method to investigate the extent of mutagenesis at specific target sites. Thus, robust and simple-to-use software that enables researchers to retrieve mutation profiles from NGS data is needed. Here, we present Sequence Interrogation and Quantification (SIQ), a tool that can analyse sequence data of any targeted experiment (e.g. CRISPR, I-SceI, TALENs) with a focus on event classification such as deletions, single-nucleotide variations, (templated) insertions and tandem duplications. SIQ results can be directly analysed and visualized via SIQPlotteR, an interactive web tool that we made freely available. Using novel and insightful tornado plot visualizations as outputs we illustrate that SIQ readily identifies differences in mutational signatures obtained from various DNA-repair deficient genetic backgrounds. SIQ greatly facilitates the interpretation of complex sequence data by establishing mutational profiles at specific loci and is, to our knowledge, the first tool that can analyse Sanger sequence data as well as short and long-read NGS data (e.g. Illumina and PacBio).

INTRODUCTION

The broad implementation of CRISPR technology in biological and biomedical research has led to an expansion of approaches that rely on robust and correct interpretation of sequence changes that result from repair of single-strand or double-strand DNA breaks. All outcomes combined thus represent the repair profile of a particular DNA damage in a particular cellular or organismal context. Upon experimental perturbations, deep-sequencing of PCR amplicons using NGS techniques has become a valuable tool to obtain detailed information of the underlying mutagenic mechanisms. Yet, not all researchers are able to analyse these data as there is a lack of tools that are designed to be used by researchers without training in informatics. To meet this increasingly growing demand, we designed and created SIQ, for Sequence Interrogation and Quantification. SIQ, which can be run on any computer system and uses the raw sequencing files as input to classify and quantify the identified sequence variants. It can run multiple files simultaneously and the resulting Excel file can be data-mined but also uploaded in SIQPlotteR, an interactive web tool we designed that allows for extensive data visualization and exploration (https://siq.researchlumc.nl/SIQPlotteR/).

RESULTS

SIQ method

SIQ utilizes data obtained by sequencing PCR products covering a target site of interest that, for instance, has been targeted *in vivo* by a nuclease and subsequently has been repaired by cellular repair pathways (Figure 1A). It can process collections of capillary (Sanger) sequences, each containing a single mutation, but the true strength of SIQ is that it can identify mutational profiles in pooled DNA containing an extensive mix of mutational outcomes and deep sequenced by NGS methods (Figure S1). For SIQ-analysis of experiments where short-read sequencing will be applied (i.e. Illumina paired-end sequencing) we recommend a PCR amplicon of <290bp (for 2x150bp paired-end reads) or <580bp (for 2x300bp paired-end reads) to ensure that the reads contain some overlap. For long-read sequencing these criteria do not apply and larger amplicons can be used (e.g. >3kb).

NGS data can directly be analysed by SIQ, which includes a graphical user interface, designed to run on any operating system (Windows, Linux, MacOS, Figure S2). SIQ can also run from the command line, if desired. Apart from sequence data, SIQ requires a reference DNA sequence as input. What sets SIQ apart from other mapping approaches is that it is specifically designed to detect sequence changes at an expected target site (e.g. a CRISPR/Cas9, Cas12, I-SceI, TALEN, base-editor or AsiSi site) to focus on identifying variants at or in close proximity to the expected target site. This is achieved by performing a k-mer mapping strategy which detects matching sequences flanking the target site. If paired-end reads are used as input, they are first merged into a single read (via FLASH2 (*1*)). Reads are then passed through various filters, which includes removing low-quality and non-informative bases. Reads that pass the filters are mapped to the reference DNA file (Figure 1A). Subsequently, event classification is carried out using categories that reflect common genetic variations, such as deletion, insertion, single-nucleotide variations (SNVs) and wild-type. Additional classification concerns deletions that also contain unmatched bases in between the deletion junctions, hence reflecting insertions. For cases in which (part of) the insertion can be reliably matched (surviving statistical scrutiny) to DNA sequences surrounding the mutation, the event is classified as a templated insertion. Such templated insertions have been previously shown to be a unique hallmark of the double-strand break repair pathway theta-mediated end-joining (TMEJ) (*2*). Another classification that SIQ reports are tandem duplications: an insertion (≥6bp) that is an exact duplicate of the DNA sequence immediately flanking it. Especially Cas9 nickase enzymes, combined with two sgRNAs targeting opposite strands, produce DNA breaks with protruding ends resulting in this type of genetic alterations (*3-6*). Finally, SIQ is also able to detect precise gene editing outcomes by matching the reads to a supplied repair template; such events are classified as homology-directed repair (HDR). In addition to event classification, additional metadata such as event location with respect to cut site, size, and micro-homology usage are determined. SIQ can process multiple sequence files and targets simultaneously: even on a regular computer SIQ can analyse millions of reads per minute.

The output of SIQ is an Excel table, which can be analysed directly, or be processed through a dedicated web-tool called SIQPlotteR, which we made publicly available: https://siq.researchlumc.nl/SIQPlotteR/; SIQPlotteR can also be installed locally.
We created SIQPlotter as we experienced that the amount of data produced by targeted sequence experiments requires condensed data visualizations as well as an interactive environment to allow researchers to explore the data from different angles. To capture the entire spectrum of mutational outcomes we developed a novel visualisation that we termed ‘tornado plot’, which shows in a single graph the repair outcome type, the weight of each mutation, the extent of micro-homology at the junctions, and the location of the event with respect to the target site (Figure 1B).

Mutation analysis on cells treated with CRISPR-Cas9 variants

To showcase SIQ, we processed a series of experiments. We induced double-strand breaks (DSBs) with different configurations in mouse ES (mES) cells: blunt DSBs using wild-type Cas9, and DSBs with 5’ and 3’ protruding overhangs using Cas9 nickases (Cas9D10A and Cas9N863A, respectively). In addition to wild-type mES cells we included cells with a deficiency in Pol⍰, which is critically important for TMEJ, and cells with a deficiency in Ku80, which is a key factor in non-homologous end-joining (NHEJ). To generate data sets rich with different variants we targeted the selectable gene *Hprt*, which when mutated, confers resistance to treatment with 6-thio guanine (6-TG). DNA from 6-TG-resistant cells was isolated for each cell line and amplified using specific primers (Supplemental Table 1). Deep sequencing was performed on these amplicons and the data was analysed by SIQ and subsequently visualized using SIQplotteR (Figure 2). Importantly, cellular selection is not a prerequisite for establishing detailed mutation profiles as NGS data from pools of unselected cells also produce a wide spectrum of mutational outcomes, even if they contain up to 90% wild-type reads, which can be filtered out in SIQplotteR (Figure 1B).

Figure 2 demonstrates that SIQPlotteR visualizes SIQ-processed NGS data in intuitively interpretable formats, which can be adapted in several dimensions (types of outcome to visualize; scales; colour coding; sorting, Figure S3). Furthermore, the data obtained with SIQ-analysis recapitulates the repair profiles that have been previously found for the tested conditions. Cas9 WT and Cas9 nickase variants produce entirely different mutation profiles: Cas9 WT-induced blunt DSBs creates small deletions that, in mES cells, are characterized by micro-homology and 1bp insertions; Cas9 nickase-induced DSBs with 3' overhangs (Cas9N863A) predominantly give rise to tandem duplications, whereas DSBs with 5' overhangs (Cas9D10A) mostly produce deletions in which the 5' protruding sequence has been lost, but also tandem duplications, in which fill-in has occurred (Figure 2A-C, Figure S4) (*6, 7*). Overtly different mutation profiles are produced in cells that contain DNA-repair deficiencies, such as in TMEJ and NHEJ deficient cells. Confirming published work, Figure 2 shows a prominent role for Pol⍰ in mutagenic repair of DSBs induced by Cas9 WT, leading to a characteristic micro-homology-mediated repair profile (i.e. the two blue blocks, Figure 2A, Figure 2E) (*6, 8*). The action of NHEJ can also be observed in wild-type cells, as it is reflected by the presence of 1 bp insertions (Figure 2A and 2D, purple) that are KU80 dependent. In addition, Figure 2D further highlights the following genetic requirements: i) a Pol⍰ dependency for deletions containing templated insertions, which are increasingly manifest in *Ku80*$^{-/-}$ cells, ii) a Ku80 dependency for tandem duplications at DBS having 5' protruding ends, and iii) also a Ku80 dependency for tandem duplications at DBS having 3' protruding ends.

To illustrate the ability of SIQ to determine mutation spectra that are the consequence of repair of other types of DNA damage than nuclease-induced DBSs we analysed two data sets. First, we used data derived from *C. elegans* FancJ mutants in which DSBs spontaneously occur at G-rich sequences, as these can form stable DNA secondary structures called G-quadruplexes (G4s), which impede ongoing DNA replication (*9*). In the absence of FANCJ/DOG-1 in *C. elegans* genomic deletions arise that have lost a G4 motif as well as 50-200 bases of downstream sequence (*10-12*). Performing SIQ on NGS data of targeted sequencing around such G4 sites in worm populations produce G4 deletion spectra that recapitulates repair of G4-induced DSBs at an unprecedented scale (Figure 2F). Second, we used published data from another research group that used base-editing CRISPR-technology to induce specific base-substitutions at the EMX1 locus in HEK293T cells (*13*). Figure 2G shows the output of SIQplotteR that visualises the presence of base alterations at a given target, in this case the result of base-editing EMX1 in HEK293T cells, which are dominated by mutations to TT at the target site (Figure 2G).

SIQ on long-read PacBio data

While short read sequencing (e.g. by illumina platforms) is often informative and affordable, the use of long-read sequencing starts to gain momentum as it allows for inclusion of large structural variants in the analysis. Yet also for their output, easy processing tools for user-friendly quantification and inspection are grosso modo missing. Therefore, we designed the current version of SIQ to also create mutation profiles from long-read NGS data (i.e. PacBio data). To generate proof of concept we isolated DNA from cells that were transfected with Cas9WT and either of two different sgRNAs that induce DSBs in exon 3 of the *Hprt* gene. We designed primers to produce amplicons of 270bp and 3kb (Figure 3A) to compare short and long-read sequencing. PCR products were obtained and sequenced on a PacBio sequelII and on an Illumina HiSeq. In the size range covered by both technologies (0-200bp), we find SIQ to produce comparable spectra (Figure S4C-E). Using PacBio sequencing we can detect mutations that otherwise would be missed in Illumina sequencing as those events remove either one or both of the primers used for amplification, which constituted 8% and 12% of events, respectively (Figure 3B and 3D). In terms of mutation types (Figure 3C) and homology at the junctions (Figure 3E) the two sequence methods generate comparable footprints, with the exception of deletions with insertions (delins), which are more frequently found in PacBio data (Figure 3C). While the additional mutations detected in PacBio versus Illumina sequencing on these CRISPR sites in mES cells may appear relatively modest, research has shown that such large deletions may occur more frequently in certain cell types and species and that long-read sequencing provides a powerful method to detect undesired genome modifications (*14*).

DISCUSSION

Advantages of SIQ

Here, we have developed user-friendly software to translate complex NGS outcomes into an Excel file format that allows for multifactorial data-mining, and into intuitive and easily amendable graphics to facilitate interpretation. Because SIQ only needs NGS output and a reference target that is suspected of having mutations, it can be used to create mutation profiles for a wide range of experimental approaches which apart from the now common CRISPR/Cas9 technology include targeting by base editors, TALENs, endonucleases, (plasmid based-)DNA crosslinks and replication blocks (e.g. via G4 quadruplexes).

SIQ is designed to facilitate researchers that do not have in-depth knowledge on how to handle NGS data, or are not skilled in programming: a user can simply select the amplicon NGS files to be analysed and input the DNA reference. Once the target location(s) are set and the primers used are added (optionally), analysis can commence. When analysis is complete the resulting Excel table can be data-mined via the numerous parameters that are annotated to each mutational outcome. The Excel table can also be directly uploaded in SIQPlotteR to analyse data quality as well as to generate various interactive data visualizations. We have included visualizations that show: quality control, targeting frequency, repair-type classification, size alteration, micro-homology, SNV alteration and the insightful tornado plots. For all of these plots we allow users to filter experiments based on the number of reads or event type, select and sort samples, choose colours and to finally export their plots to a PDF format. To be able to test the utility of SIQ we uploaded data used in this manuscript to be directly tried in SIQPlotteR (https://siq.researchlumc.nl/SIQPlotteR).

Comparison to other methods

In recent years several tools have been created to analyse amplicon data, such as Amplican (*15*), Crispresso2 (*16*), CrispRvariants (*17*), ScarMapper (*18*) and CRISPAltRations (*19*). In general, these tools have been designed to analyse a specific type of CRISPR editing. Some tools, such as CRISPAltRations are specifically trained to detect CRISPR edits in a limited window around the break site, precluding detection of other types of events, such as large deletions, or its use in analysing experiments that employ other means of creating DNA alterations. While most tools provide basic classification of events, such as deletions and insertions, none of these report tandem duplications or templated insertions.

Apart from generating multi-dimensional output visualizations, that can easily be modified, another major difference to the now available tools is the ease of installation and usage of SIQ. Some of the current tools require additional software dependencies to be installed, or cannot be run from Windows. We feel that in most cases, (bio)informatic expertise is needed or nearby experts to install the software and run analyses. To optimally facilitate unrestricted data processing, without restrictions on accessibility, file size, number limits, we developed SIQ to not depend on websites, but instead operate on a local computer and to implement it in Java to allow researchers to simply launch SIQ upon download.

MATERIAL AND METHODS

Cell culture and transfection

129/Ola-derived IB10 mouse embryonic stem (mES) cells were cultured on gelatin-coated plates in Buffalo rat liver (BRL)-conditioned mES cell medium (Dulbecco’s modified Eagle’s medium (Gibco) supplemented with 100 U/ml penicillin, 100µg/ml streptomycin, 2mM GlutaMAX, 1mM sodium pyruvate, 1× non-essential amino acids, 100µM β-mercaptoethanol (all from Gibco), 10% fetal calf serum and leukemia inhibitory factor). *HPRT-eGFP* wild-type, *Polq*$^{-/-}$ and *Ku80*$^{-/-}$ mES cells were generated as previously described (*7*). Cells were transfected in suspension using a lipofectamine 2000 (Invitrogen):DNA ratio of 2.4:1. Briefly, 1.5×10^{6} cells were transfected using 3µg of total DNA and incubated for 30min at 37°C and 5% CO$_2$ in round-bottom tubes, subsequently cells were seeded on gelatin-coated plates containing BRL-conditioned medium.

HPRT-targeting assay

spSpCas9(BB)-2A-GFP (a gift from Feng Zhang, Addgene plasmid #48138), pU6-(BbsI)_CBh-Cas9-T2A-mCherry (a gift from Ralf Kuehn, Addgene plasmid #64324) and CBh-Cas9-Nickase-T2A-mCherry constructs containing sgRNAs were used to transfect mES cells (*7*). One day after transfection, the medium was refreshed. Cells were subcultured and *HPRT*-mutant cells were selected 7 days post-transfection by either sorting ≥100,000 GFP-negative cells on a BD FACSAria III (using BD FACSDiva software version 9.0.1, BD Biosciences) or by seeding 500,000 cells in 6-thioguanine containing medium, subsequently cells were allowed to grow for 5–7 days.

Targeted sequencing of Cas9-induced repair outcomes

Samples for short-read (Illumina) sequencing were prepared essentially as described before (*7*). Briefly, genomic DNA was isolated and primers specific for the targeted regions were selected (Supplementary Table 1) that yield a ~150-200 bp product on wild-type alleles and that contain adaptors for the p5 and p7 index primers (5’- GATGTGTATAAGAGACAG-3’ and 5’-CGTGTGCTCTTCCGATCT-3’ respectively). These primers were used to amplify the targeted region, PCR products were subsequently purified using AMPure XP beads (Beckman Coulter) according to the manufacturer protocol and DNA was eluted in 20 µl MQ. Flow-cell adaptor sequences were added by performing PCRs with 5 µl purified PCR-product and 0.3 µM of p5 and p7 index primers. The PCR products were purified with AMPure XP beads and eluted in 20 µl MQ. PCR samples were pooled at equimolar concentrations per target-specific PCR. The quality and quantity of these pools were analysed using a High Sensitivity DNA chip on a Bioanalyzer (Agilent) which was used to generate an equimolar library that was sequenced on a NovaSeq6000 or HiSeq4000 (Illumina) by 150-bp paired-end sequencing.

For PacBio sequencing, 5' Amino Modifier C6 (5AmMC6) modified primers (IDT) were designed (Supplementary Table 1) to yield a ~3500 bp product on wild-type alleles and that are tailed with universal sequences (5’-5AmMC6/GCAGTCGAACATGTAGCTGACTCAGGTCAC/Forward_sequence-3’ and 5'-5AmMC6/TGGATC-ACTTGTGCAAGCATCACATCGTAG/Reverse_sequence-3'). These primers were used to amplify the targeted region in 25 µl reactions using the PrimeSTAR GXL kit (Takara) and the following conditions: 98 °C for 30 s, 20 cycles of 95 °C for 15 s, 60 °C for 15 s and 68 °C for 4 min, and the final extension 68 °C for 7 min. Next, 2.5 – 3.5 ng round-one PCR product and Barcoded Universal Primers were used in a second-round PCR with PrimeSTAR GXL and the following conditions: 98 °C for 30 s, 20 cycles of 95 °C for 15 s, 64 °C for 15 s and 68 °C for 4 min, and the final extension 68 °C for 7 min. DNA concentrations were measured using the Quant-iT dsDNA assay kit and the Qubit Fluorometer (both Thermo Fisher Scientific) according to the manufacturer's protocol and PCR samples were pooled at equimolar concentrations to contain 1000-2000ng of DNA in total, the quality of these pools were analysed on the Femto pulse system (Agilent). SMRTbell library preparation was performed on 1000 ng purified PCR pool following the Procedure & Checklist - Amplicon Template Preparation and Sequencing (PN 100-815-000 Version 04, Pacific Biosciences) and using SMRTbell Express Template Prep Kit 2.0. The library was sequenced on SequelII using sequencing primer V4, Sequencing kit 2.0 and Binding kit 2.0 on an 8M SMRT cell with a movie time of 30 hr. Circular consensus sequences were generated with ccs version 6.0.0 (commit v6.0.0-2-gf165cc26) and barcodes were demultiplexed using lima 2.0.0 (commit v2.0.0).

SIQ implementation

SIQ is implemented in Java to be run on any operating system and requires at least Java 8 to run. As an initial check SIQ checks if all files can be located. In addition the user can (strongly recommended) define flanks, which define the expected target site (e.g. a CRISPR cut site). The middle between the left and right flank defines the expected target site and that location is set to 0. The provided flanks should be ≥15bp and are required to be present in the reference sequence. For target where two targets are used (e.g. if two sgRNAs are used) the flanks can be separated: the end of the left flank defines one target site and the start of the right flank defines the second target site. The primer used to perform the experiment can also be supplied (recommended) and need to be present in the reference DNA sequence as well. The primer sequences are used to ensure reads start within the defined primers. If both R1 and R2 NGS files are supplied, SIQ attempts to merge the paired-end data using Flash (v2.2.00) (*1*). SIQ then uses the merged file (or only the file in R1 if that was supplied) to map. For short-read data it will initially check the orientation of the reads and assume the same orientation is used throughout. For PacBio data the reads are used in both forward or reverse complement orientation, depending on the read. Bases below a base quality threshold are cut off, leaving a high quality read. That read is then mapped to the supplied reference. For short-read sequencing the read should start within the primer binding sites and the detected event should start at least 5 bases (optional and configurable) away from the primer binding sites. This ensures mutagenic events are only detected if the primers annealed at the intended location in the DNA. SIQ classifies the reads based on the difference with the provided reference sequence and outputs an Excel table.

Templated insertions are insertions that are copied from a nearby stretch of DNA. To determine if a delins (deletion with insertion) is a templated insertion the inserted sequence is searched around the deletion junction. This is only performed for delins with an insertion of ≥6bp as it is not possible for smaller insertions to determine the origin of the insertion (random chance of finding that sequence is too high). The search space is predefined to 100bp (configurable) up- and downstream of the left junction (start point of the deletion) and the right junction (endpoint of the deletion) in both forward and reverse complement orientation and selects the largest overlapping sequence with the insert. A test is then performed to ensure that the probability of finding such a match is <10% when the junctions are shuffled. So only if an insertion with a large enough match in the flank is found it is classified as a templated insertion (tins).

Tandem duplications are insertions that exactly match the left or right junction and are ≥6 nucleotides long. Tandem duplication compound (TD+) are insertions where part of the insertion exactly matches the left or right junction.

SIQPlotteR implementation

SIQPlotteR code was written using R (https://www.r-project.org) and Rstudio (https://www.rstudio.com). To run the app, several freely available packages are required: shiny, ggplot2, dplyr, lobstr, colourpicker, grid, gridExtra, readxl, shinyWidgets, tidyr, RColorBrewer, sortable, ggpubr, ggrepel, DT, gplots, FactoMineR, factoextra, umap. Up-to-date code and new releases will be made available on GitHub, together with information on running the shiny app locally: https://github.com/RobinVanSchendel/SIQ.

The GitHub page of SIQ is the preferred way to communicate issues and request features (https://github.com/RobinVanSchendel/SIQ/issues). Alternatively, users can contact the developers by e-mail or Twitter. Contact information is found on the GitHub page. SIQPlotteR can be installed locally or you can use the available website to analyse SIQ output: https://siq.researchlumc.nl/SIQPlotteR/

Generation of in silico datasets

To generate in silico datasets we generated 267 datasets based on target sites in the human genome. For each dataset set we created 11 subsets with a variable mutation frequency ranging from 0 – 1 with 0.1 step size. For each subset we created 10,000 reads that were either wild-type or contained deletions and insertions ranging from -25 to +25 bp. To introduce sequencing errors into our sets we ran ART (*20*) to obtain a set with sequencing errors (options used: -na -ss HSXn -qs 10 -qs2 10). These sets were subsequently analysed by SIQ, Crispresso2 (*16*)and Amplican (*15*).

C. elegans G4 experiment

A single animal of the strain XF1520 with genotype *dog-1(gk10)* was put on a 6cm dish containing NGM and OP50. One week after plating the plate was full and animals were rinsed off in MQ, washed five times with MQ, and DNA was isolated using a DNA blood & tissue culture kit (Qiagen) following the manufacturer's protocol. 1µl of DNA was PCRed using primers at the G4 site qua2478 (see Supplemental Table 1) and processed as described above to generate an NGS library.

AVAILABILITY

The latest version of SIQ and SIQPlotteR are available in the GitHub repository (https://github.com/RobinVanSchendel/SIQ). Since this software is commonly used in our lab we expect to develop and extend it further in the future.

ACCESSION NUMBERS

The raw targeted sequencing data generated in this study has been deposited in the NCBI SRA database under accession PRJNA802705. The base editor data for EMX1 used for Figure 2G was downloaded from accession number SRR3305545. CAS9D10A data was previously generated and can be found in accession numbers: SRR12079930, SRR12079938 and SRR12079923, and Cas9N863A in WT cells in accession number SRR12079956.

SUPPLEMENTARY DATA

Supplementary Table 1 – Primer sequences used for NGS

Supplementary Table 2 – sgRNA sequences used in this study

ACKNOWLEDGEMENT

We thank members of the Tijsterman lab for critical testing of SIQ and SIQPlotteR and for critical reading of the manuscript. We thank Dr. Bert van de Kooij for critical reading of the manuscript.

FUNDING

J.S. is supported by a Young Investigator Grant from the Dutch Cancer Society (KWF, 2020-1/12925); M.T. is supported by grants from the Dutch Cancer Society (11251/2017-2) and the Holland Proton Therapy Centre (2019020-PROTON-DDR) and by an ALW OPEN grant (OP.393) from the Netherlands Organization for Scientific Research for Earth and Life Sciences.

TABLE AND FIGURE LEGENDS

Figure 1*. Implementation of SIQ*

A) Schematic illustration of SIQ. NGS or Sanger sequencing on a PCR amplicon is required as input for SIQ, together with a reference DNA fasta file. Reads are optionally merged and SIQ produces an Excel output table, which can also directly be analysed by SIQPlotteR. B) Examples of SIQPlotteR tornado plot visualisations for a target site in the mouse HPRT gene exon 2. Each colour represents an event type and the height of each colour represents the contribution to the total fraction. The white space represents the deletion size for each event. For deletions additional colouring is added, based on the extent of micro-homology found at the junctions or the presence of insertions. Left panel shows all events, right panel excludes the wild-type events.

Figure 2*. Showcasing SIQ capabilities*

A) Tornado plot representation of mutation profiles at Cas9-induced DSB in mES cells of the indicated genotypes, which were transfected with Cas9WT and HPRT_Ex3.1 sgRNA. Data is plotted relative to the Cas9WT break site (at 0) and sorted by deletion size. Tins (templated insertions, see Methods), 1bp insertions, insertions and deletion insertions (delins) are colour-coded. The degree of blue colouring reflects the extent of micro-homology found at the deletion junction. B) Similar to A, but for WT cells transfected with Cas9D10A in combination with two sgRNAs in opposite strands of the DNA to create a break with a 5' overhang of 50bp. Td and td+ (tandem duplication and tandem duplication compound, see Methods) are depicted in orange and indicate the DNA that has been duplicated. C) Similar to A, but for WT cells transfected with Cas9N863A in combination with two sgRNAs in opposite strand of the DNA to create a break with a 3' overhang of 43bp. D) Fraction of mutation types identified for each Cas9 variant in mES cells with indicated genotypes. E) The degree of micro-homology that is found at the junctions of deletions and tandem duplications for each Cas9 variants in the mES cells with the indicated genotype. F) Tornado plot (inverted) representing the mutation profile at a G4 site in DNA extracted from DOG-1 deficient *C. elegans*. The PCR strategy chosen generates a ~350bp PCR product, which excludes wild-type events from being detected (reads are 2x150bp). Deletion products are ordered on end position. G) A SNV-alteration plot that displays base-editing at an EMX1 site in data obtained from (*13*). The x-axis displays the location relative to the target site and the y-axis the fraction of total reads. Dinucleotide SNVs are detected as delins (del = 2, ins = 2) and can be optionally displayed in SIQPlotteR.

Figure 3*. SIQ on PacBio data*

A) PCR strategy for Illumina and PacBio sequencing of the mouse HPRT gene targeted at exon 3 with either HPRT_Ex3.1 or Ex3.2 sgRNA. B) Tornado plot representation of mutation profiles at Cas9-induced DSB in mES cells, which were transfected with Cas9WT and HPRT_Ex3.1 sgRNA and sequenced with Illumina or PacBio. Data is plotted relative to the Cas9WT break site (at 0) and sorted by deletion size. Tins (templated insertions, see Methods), 1bp insertions, insertions and deletion insertions (delins) are colour-coded. The degree of blue colouring reflects the extent of micro-homology found at the deletion junction. C) Fraction of mutation types identified for HPRT_Ex3.1 and HPRT_Ex3.2 for the indicated sequencing strategy. D) As in B, but now for HPRT_Ex3.2. E) The degree of micro-homology that is found at the junctions of deletions for HPRT_Ex3.1 and HPRT_Ex3.2 for the indicated sequencing strategy.

REFERENCES

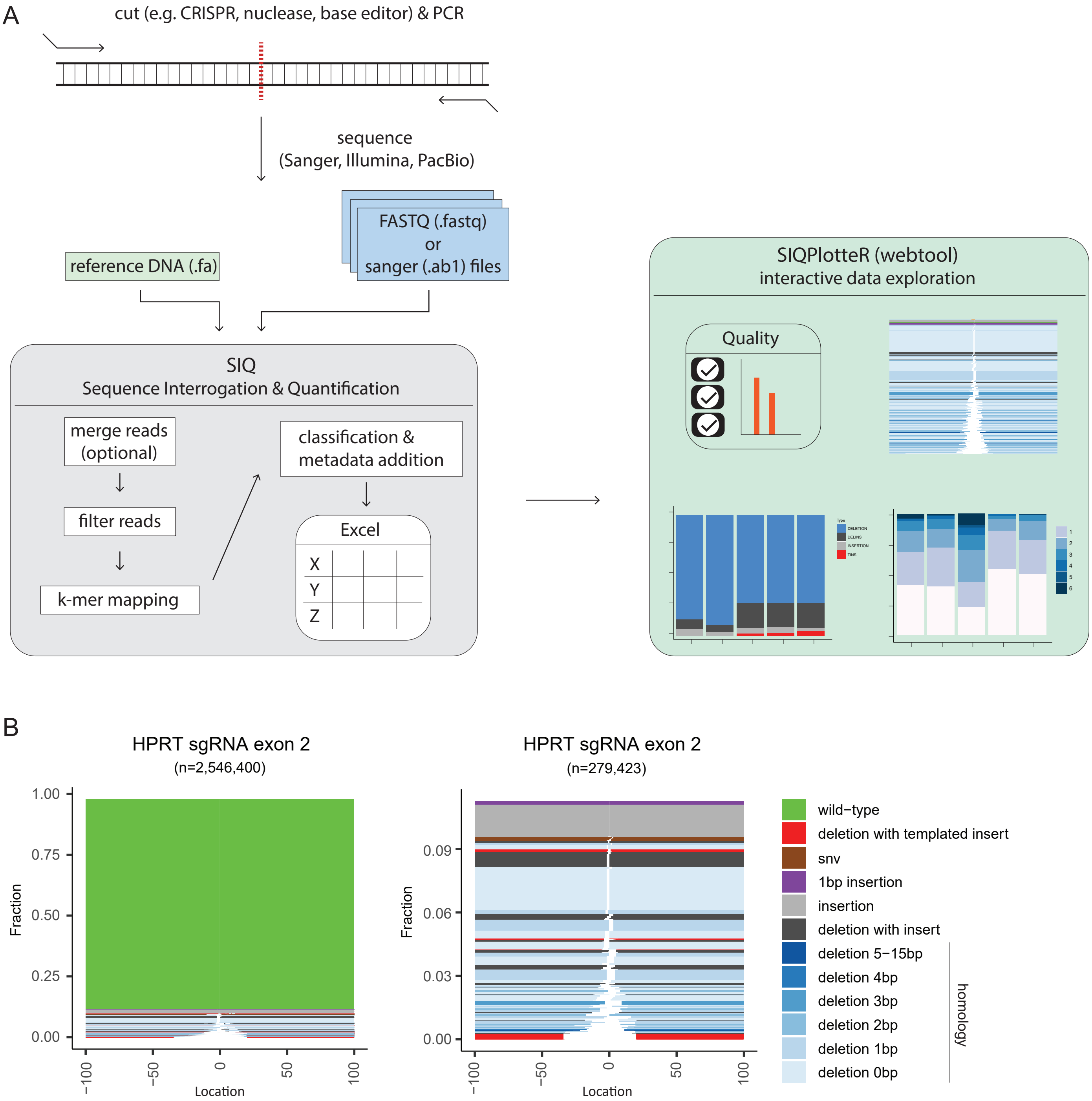
1. T. Magoč, S. L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963 (2011).
2. J. Schimmel, R. van Schendel, J. T. den Dunnen, M. Tijsterman, Templated Insertions: A Smoking Gun for Polymerase Theta-Mediated End Joining. *Trends Genet* **35**, 632-644 (2019).
3. A. Bothmer *et al.*, Characterization of the interplay between DNA repair and CRISPR/Cas9-induced DNA lesions at an endogenous locus. *Nature Communications* **8**, 13905 (2017).
4. P. J. Chen *et al.*, Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. *Cell* **184**, 5635-5652.e5629 (2021).
5. S. Schiml, F. Fauser, H. Puchta, Repair of adjacent single-strand breaks is often accompanied by the formation of tandem sequence duplications in plant genomes. *Proceedings of the National Academy of Sciences* **113**, 7266-7271 (2016).
6. J. Schimmel, H. Kool, R. van Schendel, M. Tijsterman, Mutational signatures of non-homologous and polymerase theta-mediated end-joining in embryonic stem cells. *The EMBO Journal* **36**, 3634-3649 (2017).
7. J. Schimmel, N. Muñoz-Subirana, H. Kool, R. van Schendel, M. Tijsterman, Small tandem DNA duplications result from CST-guided Pol α-primase action at DNA break termini. *Nat Commun* **12**, 4843 (2021).
8. D. W. Wyatt *et al.*, Essential Roles for Polymerase θ-Mediated End Joining in the Repair of Chromosome Breaks. *Mol Cell* **63**, 662-673 (2016).
9. I. Cheung, M. Schertzer, A. Rose, P. M. Lansdorp, Disruption of dog-1 in Caenorhabditis elegans triggers deletions upstream of guanine-rich DNA. *Nat Genet* **31**, 405-409 (2002).
10. W. Koole *et al.*, A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. *Nature Communications* **5**, 3216 (2014).
11. E. Kruisselbrink *et al.*, Mutagenic capacity of endogenous G4 DNA underlies genome instability in FANCJ-defective C. elegans. *Curr Biol* **18**, 900-905 (2008).
12. R. v. Schendel, R. Romeijn, H. Buijs, M. Tijsterman, Preservation of lagging strand integrity at sites of stalled replication by Pol &#x3b1;-primase and 9-1-1 complex. *Science Advances* **7**, eabf2278 (2021).
13. A. C. Komor, Y. B. Kim, M. S. Packer, J. A. Zuris, D. R. Liu, Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420-424 (2016).
14. I. Höijer *et al.*, CRISPR-Cas9 induces large structural variants at on-target and off-target sites in vivo that segregate across generations. *Nature Communications* **13**, 627 (2022).
15. K. Labun *et al.*, Accurate analysis of genuine CRISPR editing events with ampliCan. *Genome Res* **29**, 843-847 (2019).
16. K. Clement *et al.*, CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat Biotechnol* **37**, 224-226 (2019).
17. H. Lindsay *et al.*, CrispRVariants charts the mutation spectrum of genome engineering experiments. *Nature Biotechnology* **34**, 701-702 (2016).

18. W. Feng *et al.*, Marker-free quantification of repair pathway utilization at Cas9-induced double-strand breaks. *Nucleic Acids Res* **49**, 5095-5105 (2021).

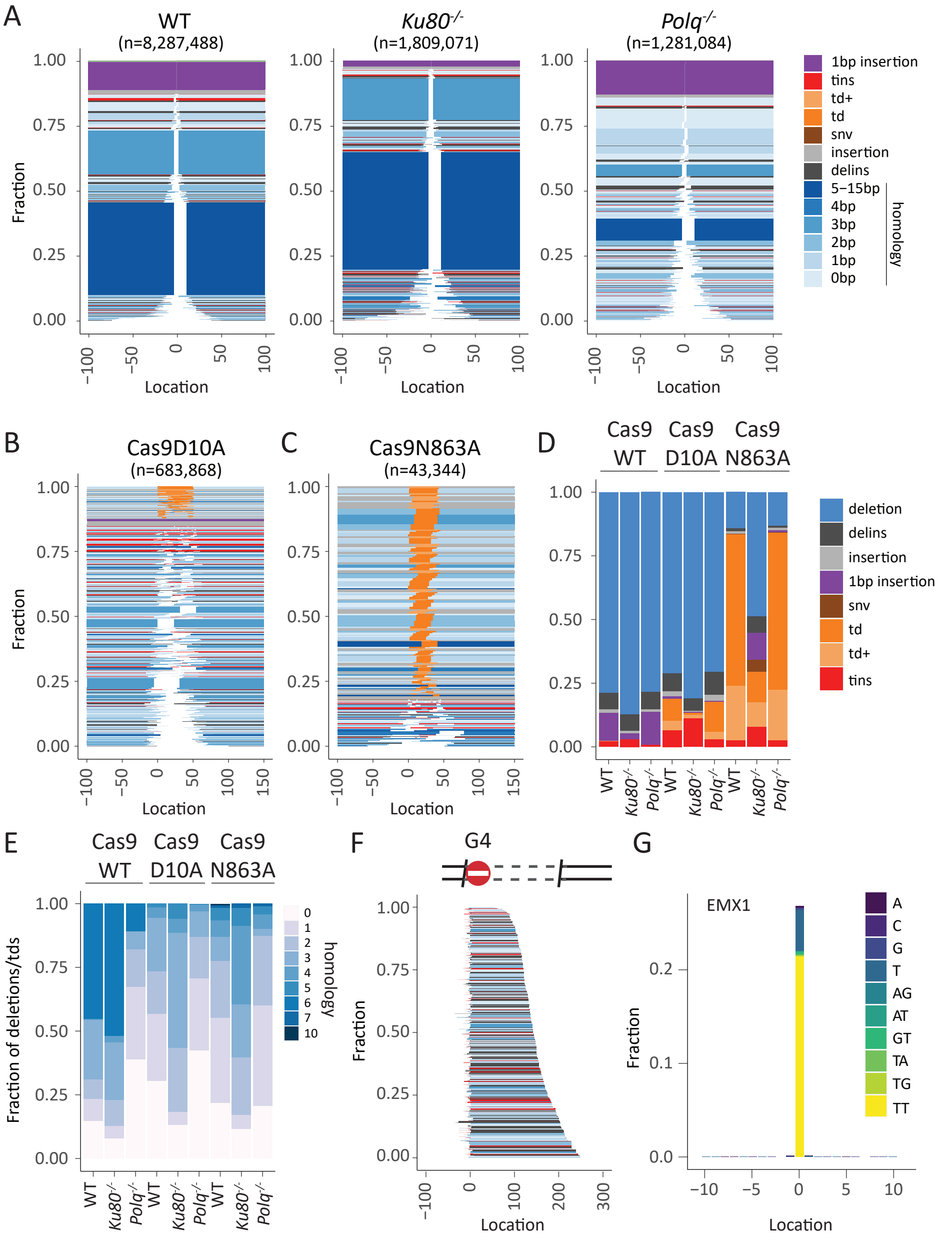
19. G. Kurgan *et al.*, CRISPAltRations: a validated cloud-based approach for interrogation of double-strand break repair mediated by CRISPR genome editing. *Mol Ther Methods Clin Dev* **21**, 478-491 (2021).

20. W. Huang, L. Li, J. R. Myers, G. T. Marth, ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593-594 (2011).

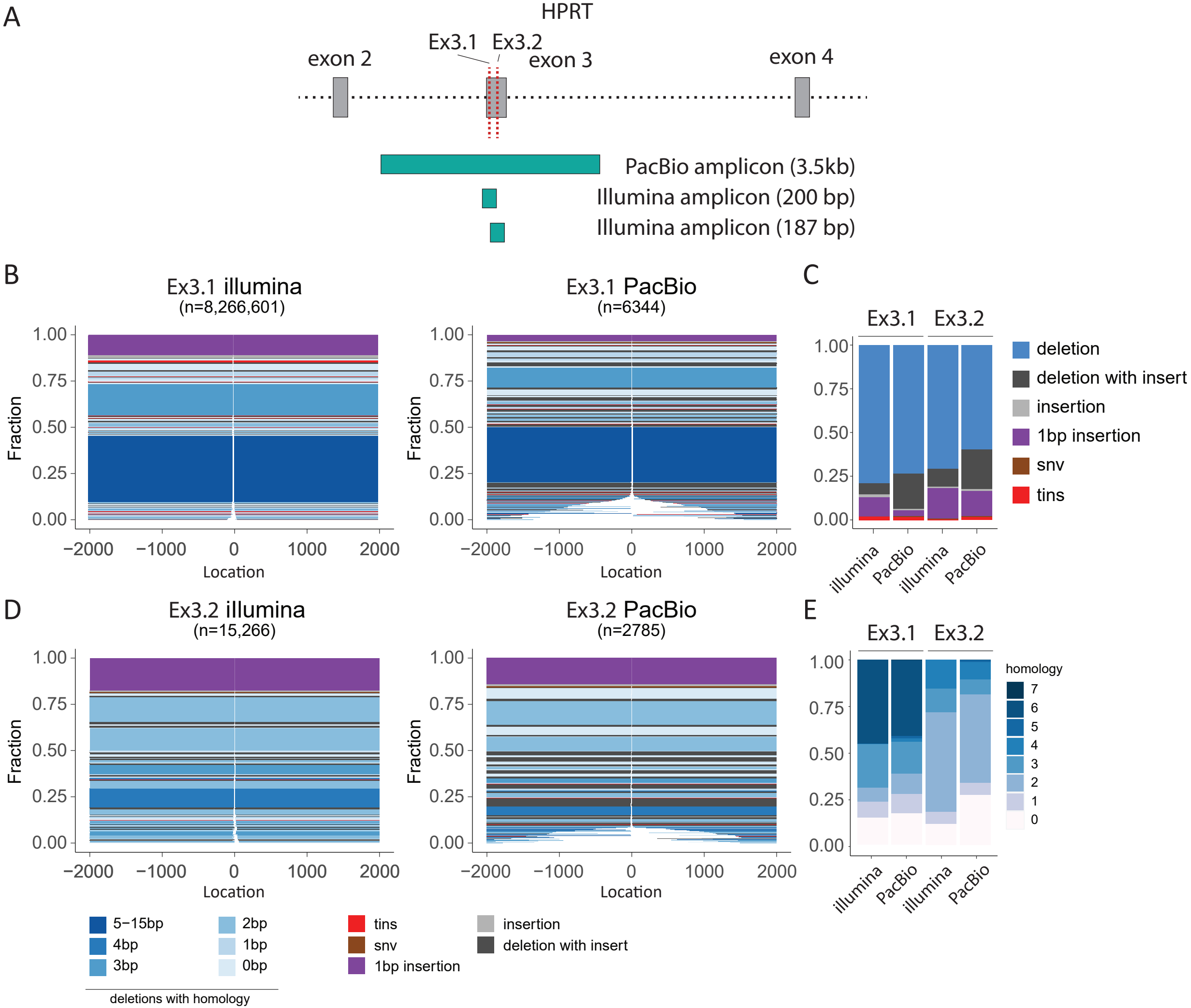
van Schendel, Figure 1



van Schendel, et al. - Figure 2



van Schendel, et al. - Figure 3


A
HPRT
Ex3.1
Ex3.2
exon 2
exon 3
exon 4
PacBio amplicon (3.5kb)
Illumina amplicon (200 bp)
Illumina amplicon (187 bp)
B
Ex3.1 illumina
(n=8,266,601)
Ex3.1 PacBio
(n=6344)
Fraction
Location
C
Ex3.1
Ex3.2
deletion
deletion with insert
insertion
1bp insertion
snv
tins
illumina
PacBio
illumina
PacBio
D
Ex3.2 illumina
(n=15,266)
Ex3.2 PacBio
(n=2785)
E
homology
5−15bp
4bp
3bp
2bp
1bp
0bp
deletions with homology
tins
snv
1bp insertion
insertion
deletion with insert